Proximity Variational Inference

Jaan Altosaar
Department of Physics
Princeton University
altosaar@princeton.edu

Rajesh Ranganath
Department of Computer Science
Princeton University
rajeshr@cs.princeton.edu

David M. Blei
Data Science Institute
Department of Computer Science and Statistics
Columbia University
david.blei@columbia.edu

Abstract

Variational inference is a method for approximating posterior distributions in latent variable models. This technique has enabled the use of Bayesian methodology in settings where it would otherwise be infeasible, yet problems remain. The optimization suffers from sensitivity to initialization and prefers underdispersed distributions. We remedy this by deriving an alternate optimization algorithm for variational inference based on proximal expansions of the variational objective with additional constraints. We derive a scalable variant that runs as fast as variational inference. In our experiments, we design an entropy constraint and show that our approach is less sensitive to initialization. We test the method in a Bernoulli factor model and a sigmoid belief net model of images trained on MNIST. In the sigmoid belief net, our model recovers good posterior predictive distributions where standard variational inference fails.

1 Introduction

Variational inference (Jordan et al., 1999) is an optimization-based inference method that tries to find the distribution in a family that is the closest in Kullback-Leibler (KL) divergence to the posterior. Despite the increase in the applicability of variational inference, problems remain. This method can suffer from bad local optima partially caused by the independence assumptions made by the posterior (Theis and Hoffman, 2015; Shah et al., 2015). Another issue occurs during optimization. After the variational approximation has removed support from a particular neighborhood, it is very hard for it to grow back (MacKay, 2003; Burda et al., 2016). The latter issue stems from the form of the KL divergence, where the constraint that p dominates q has unbounded weight. Variational inference cannot recover from bad initializations. These two problems lead to our central challenge: given a poor initialization, are there variational inference algorithms that avoid poor local optima? Can we design fast algorithms that avoid taking bad steps during variational inference?

We present *proximity variational inference (*PVI*)*, a technique for variational inference derived from changing the proximity function used in gradient ascent of the evidence lower bound (ELBO). Gradient ascent of an objective function is equivalent to minimizing the first-order Taylor expansion of the objective, subject to a proximity constraint (Spall, 2003; Boyd and Vandenberghe, 2004). Our approach builds on this, by adding additional constraints meaningful to variational methods. An example is a constraint is based on entropy. In this case, each step taken must have a similar entropy to the previous step. We demonstrate the value of this constraint in a Bernoulli factor model and sigmoid belief network model of images. Adding the entropy constraint reduces the effects of poor initialization of

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

the algorithm. The additional entropy change constraint on each update counters the local pathologies induced by the KL divergence.

Related work. KL proximal variational inference is a method for optimizing the ELBO subject to a proximal term that forces the approximate posterior to remain close on each gradient update (Khan et al., 2015). Theis and Hoffman (2015) also propose a trust-region update that constrains the variational parameters using the KL during optimization. These soft constraints are equivalent to our proposed method with a KL constraint; we allow for constraints beyond KL. Bregman divergences are used in convex optimization (Nocedal and Wright, 2006). Our method extends beyond this to nonconvex constraints such as the squared distance of the entropy, which works well in practice.

2 Variational Inference

Consider a model of data $\boldsymbol{x}$ with latent variables $\boldsymbol{z}$: $p(\boldsymbol{x}, \boldsymbol{z})$. The true posterior is $p(\boldsymbol{z}|\boldsymbol{x})$. In variational inference, the goal is minimize the KL divergence to the posterior. This is equivalent to maximizing a lower bound $\mathcal{L}$ on the evidence to arrive at a good approximate posterior distribution $q(\boldsymbol{z}; \boldsymbol{\lambda})$. The variational family q is indexed by parameters $\boldsymbol{\lambda}$ which we optimize to maximize the lower bound (Wainwright and Jordan, 2008; Hoffman et al., 2013). The ELBO is

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_q[\log p(\boldsymbol{x}, \boldsymbol{z})] - \mathbb{E}_q[\log q(\boldsymbol{\lambda})]. \tag{1}$$

The first term in this objective is the expected log-likelihood; it encourages configurations of the latent variables that maximize the likelihood of the data. The second term is the entropy of the variational distribution; it favors entropic configurations of latent variables.

3 Proximity Variational Inference

Gradient optimization of the ELBO corresponds to repeatedly optimizing a first-order Taylor approximation of the ELBO subject to a Euclidean proximity constraint. By altering the definition of proximity, we develop proximity variational inference. The new proximity constraints restrict the updates of variational parameters. We enable the practitioner to design constraints to guide the parameters away from poor local optima. These updates are efficient to compute when Taylor expanded. This inference technique is flexible as it enables a variety of functional forms for the proximity operator of the variational update.

Gradient methods with proximity operators. Gradient optimization maximizes the ELBO by repeatedly following gradients of the ELBO. This iterative procedure corresponds to repeatedly maximizing the linearized ELBO subject to a proximity constraint on the current variational parameter (Spall, 2003). Formally, let $\boldsymbol{\lambda}_t$ be the variational parameters and ρ be a constant. Then consider the update equation for $\boldsymbol{\lambda}_{t+1}$:

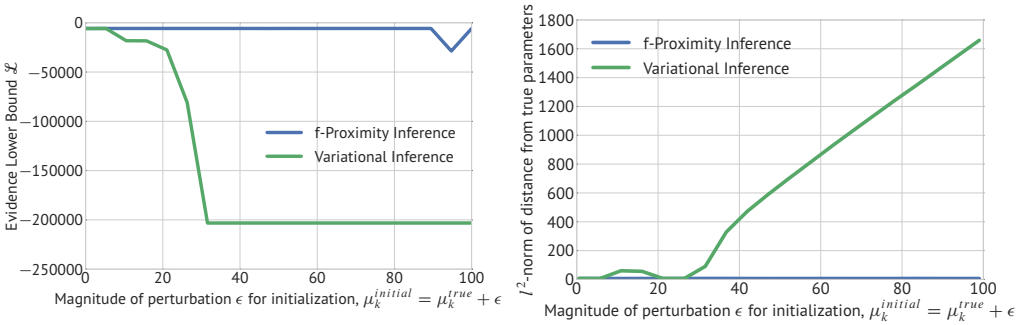
$$U(\boldsymbol{\lambda}_{t+1}) = \mathcal{L}(\boldsymbol{\lambda}_t) + \nabla\mathcal{L}(\boldsymbol{\lambda}_t)^\top(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t) - \frac{1}{2\rho}(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)^\top(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t).$$

This update equation for $\boldsymbol{\lambda}_{t+1}$ is the linearized ELBO around $\boldsymbol{\lambda}_t$ subject to $\boldsymbol{\lambda}_{t+1}$ being close in squared Euclidean distance to $\boldsymbol{\lambda}_t$. Finding the $\boldsymbol{\lambda}_{t+1}$ which maximizes U yields

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \rho\nabla\mathcal{L}(\boldsymbol{\lambda}_t). \tag{2}$$

This is equivalent to gradient ascent. This Euclidean distance-based proximity for variational inference suffers many pathologies. For example, consider a Gaussian distribution with mean zero and variance 0.01. A small change measured by Euclidean distance in the mean drastically changes where the Gaussian distribution places its support. Furthermore, the Euclidean constraint fails to prevent pathologies such as making a rapid changes in the approximation due to poor initialization. We propose enriching this class of proximity constraints based on any function f of the variational parameters. Let d be a differentiable distance function. We define the proximity update equation for the variational parameters $\boldsymbol{\lambda}_{t+1}$ to be

$$U(\boldsymbol{\lambda}_{t+1}) = \mathcal{L}(\boldsymbol{\lambda}_t) + \nabla\mathcal{L}(\boldsymbol{\lambda}_t)^\top(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t) - \frac{1}{2\rho}(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)^\top(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t) - kd(f(\boldsymbol{\lambda}_t), f(\boldsymbol{\lambda}_{t+1})). \tag{3}$$



(a) Variational lower bounds on evidence

(b) Norm of distance from truth

Figure 1: Plots of (a) the evidence lower bound and (b) normed distance from the correct cluster means for a Bernoulli spike-and-slab factor model (Section 4). The data is synthetic where the true means are known. The x-axis represents the perturbation added to the initialization of the cluster means. Initialized close to the truth, variational inference and proximity variational inference succeed. But for bad initializations, only proximity variational inference recovers the correct solutions.

This update enforces that the variational distribution be close in the statistic f. As a concrete example consider a constraint built from the entropy; if the entropy is not analytic we can estimate it using Monte Carlo. Informally, the entropy of a distribution measures the amount of randomness present in that distribution. High entropy distributions look more uniform across their support, while low entropy distributions are peaky. Formally, the entropy $H(\boldsymbol{\lambda})$ equals $-\mathbb{E}_q[\log q]$. Adding this constraint to Equation 3 as $f(\cdot) = H(\cdot)$ constraints all updates to have entropy close to their previous update. When the variational distributions are initialized with large entropy, this type of constraint is designed to balance the strong tendency toward underdispersed solutions exhibited by variational inference. This is an issue in variational inference especially if the parameters have been initialized poorly.

The update in Equation 3 rarely has a closed form solution and requires a gradient-based optimization procedure. This procedure works by perturbing the current value of λ_t to initialize λ_{t+1}, followed by gradient optimization of U. Convergence can be monitored by looking at the sign of U. When it is positive, the linearization dominates the other terms in the update equation.

Linearizing the proximity function. Equation (3) with constraints can avoid pathologies in variational inference, but it requires an internal optimization loop to compute each parameter. This is computationally burdensome. When a closed-form solution to Equation 3 is unavailable, we can use the first-order Taylor expansion of the proximity function. Letting ∇d be the gradient with respect to the second argument of the distance function and c be the first argument to the distance, we compute this expansion around $\boldsymbol{\lambda}_t$ (the variational parameters at step t):

$$U(\boldsymbol{\lambda}_{t+1}) = \mathcal{L}(\boldsymbol{\lambda}_t) + \nabla\mathcal{L}(\boldsymbol{\lambda}_t)^\top(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t) - \frac{1}{2\rho}(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)^\top(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)$$
$$- k(d(c, f(\boldsymbol{\lambda}_t)) + \nabla d(c, f(\boldsymbol{\lambda}_t))\nabla f(\boldsymbol{\lambda}_t)^\top(\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t)).$$

This linearization has closed-form update for $\boldsymbol{\lambda}_{t+1}$:

$$\boldsymbol{\lambda}_{t+1} = \lambda_t + \rho(\nabla\mathcal{L}(\lambda_t) - k(\nabla d(c, f(\boldsymbol{\lambda}_t))\nabla f(\boldsymbol{\lambda}_t)). \tag{4}$$

If the constant c is set to be the value of the proximity function at the current iterate $f(\boldsymbol{\lambda}_t)$, the added proximity has no effect. Distance functions are minimized at zero so their derivative is zero there. This means c must be set to something else. We choose the m-step lagged value of the proximity function $f(\boldsymbol{\lambda}_{t-m})$.[1] This imposes a constraint on how much the optimization can change the property f of the variational distribution over m iterations, in contrast to the standard update in Equation 2. The update in Equation (4) has the form of standard gradient ascent. It implies a global objective which varies over time:

$$\mathcal{L}_{\text{proximity}}(\boldsymbol{\lambda}_{t+1}) = \mathbb{E}_q[\log p(\boldsymbol{x}, \boldsymbol{z})] - \mathbb{E}_q[\log q(\boldsymbol{\lambda}_{t+1})] - k\,d(f(\boldsymbol{\lambda}_{t-m}), f(\boldsymbol{\lambda}_{t+1})). \tag{5}$$

As d is a distance, this remains a lower bound on the evidence, but new variational approximations

[1] Technically, the maximum of $t - m$ and 1.

remain close in f to previous iterations' distributions. The complexity of this algorithm is similar to standard variational inference: proximity variational inference corresponds to the ELBO subject to the distance constraint in f. The added complexity comes from storing values of f and computing the derivative of f; no inner optimization loop is required. The magnitude k of the constraint may need to be annealed during training.

4 Experimental Results

Bernoulli spike-and-slab model. We evaluate a Bernoulli spike-and-slab factor model. This model has a Bernoulli prior on z_{ik} and a Gaussian likelihood $x_i \sim \text{Normal}(\mu = \sum_k z_{ik}\mu_k, \sigma^2 = 1)$. To study bad initialization, we set $p = 0.01$ as the initial value of the prior. We test on synthetic data in 100 dimensions—poor initializations have much greater effects in high dimensions. The variational parameters are initialized with Gaussian noise centered at the correct mean plus a perturbation ϵ which we vary, with variance one. We optimize the objectives, Equations (1) and (5), using the RMSProp optimizer (Tieleman and Hinton, 2012) with a learning rate of 0.5. We use a five-step lagged value of the entropy, with a constraint strength of $k = 10^{12}$. We anneal the constraint strength using the following schedule: $k_t = k_{t-1}(1 - \frac{t}{T})^2$, where t is the current iteration, T is the total iterations, and $k_{t=0}$ is the initial value. This ensures the constraint decays fast enough, and was the only annealing schedule we tried. Convergence is assessed by monitoring the change in the objective and thresholding the tolerance at 10^{-10}. In Figure 1 we show that proximity variational inference recovers the correct solution even for very bad parameter initializations, very far from the correct cluster means. Without annealing of the constraint, we found that PVI still recovers correct cluster means but that the lower bound is worse.

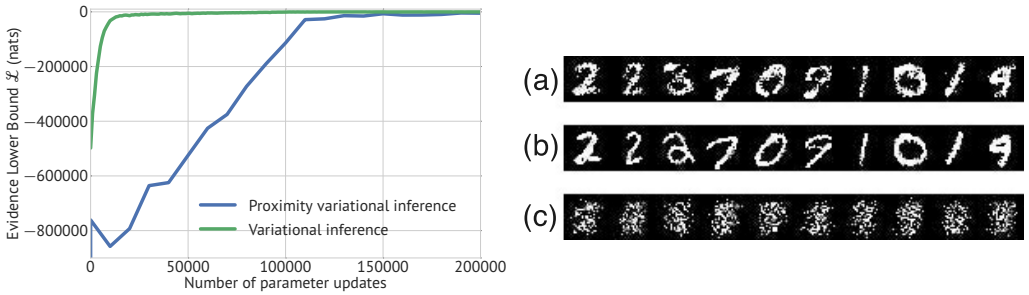


Figure 2: Poorly-initialized sigmoid belief net trained on MNIST. Left: training ELBO for the first 2×10^5 iterations. Right: samples from the posterior predictive distribution. Original MNIST digits are in the middle (b); samples for PVI and variational inference are on the top (a) and bottom (c) respectively. Proximity variational inference with the entropy constraint achieves a worse ELBO than standard variational inference. The posterior predictive samples show that the PVI constraint recovers better reconstructions of the data.

Sigmoid belief net on MNIST. We demonstrate that our method scales to large data on a common benchmark for discrete latent variables, namely the sigmoid belief net trained on binarized MNIST (Mnih and Rezende, 2016). We perturb the model parameters of a sigmoid belief net with three layers of 200 latent variables as follows. The weights and biases chosen to be poorly initialized, using Gaussian noise centered at -100 with standard deviation 0.01. We use the Adam optimizer (Kingma and Ba, 2014) with learning rate 0.001 and parameters $\beta_1 = 0.9, \beta_2 = 0.999$ and train for 3×10^6 iterations. For the control variate in (Mnih and Rezende, 2016), we take 5 samples. We use 1000 samples for estimating the ELBO. We set the prior Bernoulli parameter to 0.001, and train the model using both variational inference and proximity variational inference with the entropy constraint. For the entropy constraint, we use a magnitude of $k = 10^{10}$ and lag of 100 steps. We anneal the constraint with an exponential decay of 0.96, meaning the constraint magnitude is annealed to 0 at the end of training; this was the only schedule we tested. The ELBO was -210.0 nats and -284.8 nats for variational inference and PVI respectively. This large difference between the two methods could resolve after more iterations—the constraint magnitude k reached 0 only at the end of training. In Figure 2, we displaying samples from the posterior predictive distributions for random datapoints. While the bound for PVI is worse, it is still able to recover good reconstructions of the data. Without annealing the constraint, samples from the posterior predictive look worse but still resemble digits.

References

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2016). Importance weighted autoencoders. *ICLR preprint*, pages 1–12.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Khan, M. E., Baqué, P., Fleuret, F., and Fua, P. (2015). Kullback-leibler proximal variational inference. In *Advances in Neural Information Processing Systems*, pages 3384–3392.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Mnih, A. and Rezende, D. J. (2016). Variational inference for monte carlo objectives. *CoRR*, abs/1602.06725.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization, Second Edition*. Springer.

Shah, A., Knowles, D., and Ghahramani, Z. (2015). An empirical study of stochastic variational inference algorithms for the beta bernoulli process. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1594–1603.

Spall, J. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley.

Theis, L. and Hoffman, M. D. (2015). A trust-region method for stochastic variational inference with applications to streaming data. *ICML*.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5 - rmsprop. *COURSERA: Neural Networks for Machine Learning*.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.